Stochastic Processes

Prof. Graeme Bailey

http://cs1114.cs.cornell.edu

(notes modified from Noah Snavely, Spring 2009)

Cornell University
Computer Science

New topic: modeling sequences

- Lots of interesting things in the world can be thought of as *sequences*

- Ordering of heads/tails in multiple coin flips
- Ordering of moves in rock/paper/scissors
- Text
- Music
- Closing stock prices
- Web pages you visit on Wikipedia

How are sequences generated?

- For some sequences, each element is generated *independently*
  - Coin flips
- For others, the next element is generated deterministically
  - 1, 2, 3, 4, 5, … ?
- For others, the next element depends on previous elements, but exhibits some randomness
  - The sequence of web pages you visit on Wikipedia
  - We'll focus on these (many interesting sequences can be modeled this way)

Markov chains


Andrei Markov

- A *sequence* of random variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$

$$\boxed{\mathbf{x}_1} \rightarrow \boxed{\mathbf{x}_2} \rightarrow \boxed{\mathbf{x}_3} \rightarrow \boxed{\mathbf{x}_4} \rightarrow \boxed{\mathbf{x}_5}$$

  - $\mathbf{x}_t$ is the **state** of the model at time t
  - **Markov assumption**: each state is dependent only on the previous one
    - dependency given by a **conditional probability**:

$$p(\mathbf{x}_t | \mathbf{x}_{t-1})$$

  - This is actually a *first-order* Markov chain
  - An *N'th-order* Markov chain: $p(\mathbf{x}_t | \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-N})$

(Slide credit: Steve Seitz)

Markov chains

- Example: Springtime in Ithaca

  Three possible conditions: nice, rainy, snowy

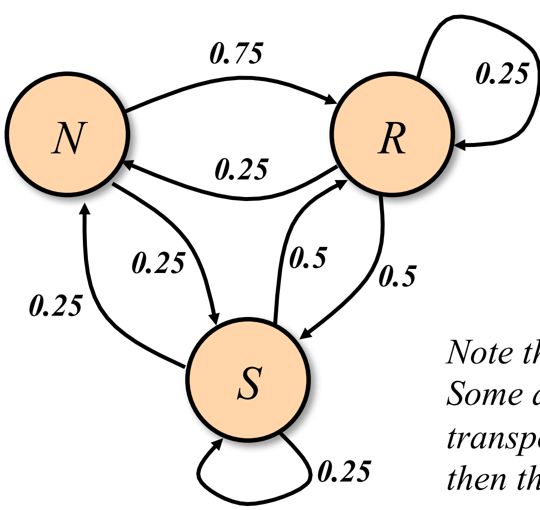
If it's nice today, then tomorrow it will be:
  rainy 75% of the time
  snowy 25% of the time

If it's rainy today, then tomorrow it will be:
  rainy 25% of the time
  nice 25% of the time
  snowy 50% of the time

If it's snowy today, then tomorrow it will be:
  rainy 50% of the time
  nice 25% of the time
  snowy 25% of the time

Markov chains

- Example: Springtime in Ithaca
- We can represent this as a kind of graph
- (N = Nice, S = Snowy, R = Rainy)



$$\mathbf{x}_{t-1} \quad \begin{array}{c} \\ \boldsymbol{N} \\ \boldsymbol{R} \\ \boldsymbol{S} \end{array} \overset{\mathbf{x}_t}{\begin{array}{ccc} \boldsymbol{N} & \boldsymbol{R} & \boldsymbol{S} \end{array}} \left[\begin{array}{ccc} 0.0 & 0.75 & 0.25 \\ 0.25 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{array} \right]$$

Transition probabilities

Note that there is no real convention for this matrix. Some authors write it this way, others prefer the transpose (ie top is 'before' and side is 'after', so then the columns sum to 1 instead of the rows).

Markov chains

- Example: Springtime in Ithaca
- We can represent this as a kind of graph
- (N = Nice, S = Snowy, R = Rainy)

$$\mathbf{x}_{t-1} \begin{array}{c} \\ \boldsymbol{N} \\ \boldsymbol{R} \\ \boldsymbol{S} \end{array} \overset{\mathbf{x}_t}{\begin{array}{ccc} \boldsymbol{N} & \boldsymbol{R} & \boldsymbol{S} \\ 0.0 & 0.75 & 0.25 \\ 0.25 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{array}}$$

Transition probabilities

If it's nice today, what's the probability that it will be nice tomorrow?

If it's nice today, what's the probability that it will be nice the day after tomorrow?

Markov chains

$$\mathbf{P} = \begin{array}{c} \boldsymbol{N} \\ \boldsymbol{R} \\ \boldsymbol{S} \end{array} \overset{\begin{array}{ccc} \boldsymbol{N} & \boldsymbol{R} & \boldsymbol{S} \end{array}}{\begin{bmatrix} 0.0 & 0.75 & 0.25 \\ 0.25 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{bmatrix}}$$

- The transition matrix at time *t*+1 is $\mathbf{P}^2$

$$\begin{bmatrix} 0.0 & 0.75 & 0.25 \\ 0.25 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{bmatrix} \begin{bmatrix} 0.0 & 0.75 & 0.25 \\ 0.25 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{bmatrix} = \begin{bmatrix} 0.2500 & 0.3125 & 0.4375 \\ 0.1875 & 0.5000 & 0.3125 \\ 0.1875 & 0.4375 & 0.3750 \end{bmatrix}$$

- The transition matrix at time *t*+n-1 $\mathbf{P}^n$

Markov chains

- What's the weather in 20 days?

$$P^{20} = \begin{bmatrix} 0.2 & 0.44 & 0.36 \\ 0.2 & 0.44 & 0.36 \\ 0.2 & 0.44 & 0.36 \end{bmatrix}$$

- Almost completely independent of the weather today
- The row [0.2 0.44 0.36] is called the stationary distribution of the Markov chain
- How might we acquire the probabilities for the transition matrix?
- One approach … 'learn' it from lots of data (eg 20 years of weather data).

Markov Chain Example: Text

"a dog is a man's best friend. it's a dog eat dog world out there."

Rows: $\mathbf{x}_{t-1}$; columns: $\mathbf{x}_t$; entries: $p(\mathbf{x}_t|\mathbf{x}_{t-1})$

	a	dog	is	man's	best	friend	it's	eat	world	out	there	.
a		2/3		1/3								
dog			1/3					1/3	1/3			
is	1											
man's					1							
best						1						
friend												1
it's	1											
eat		1										
world										1		
out											1	
there												1
.							1					

(Slide credit: Steve Seitz)

Text synthesis

- Create plausible looking poetry, love letters, term papers, etc.
- Most basic algorithm:
  1. Build transition matrix
     - find all blocks of N consecutive words/letters in training documents
     - compute probability of occurrence $p(\mathbf{x}_t | \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-(n-1)})$
  2. Given words $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}$
     - compute $\mathbf{x}_k$ by sampling from $p(\mathbf{x}_t | \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-(n-1)})$

- We can do the same sorts of things for music (notes, rhythm, harmonic structure, etc), paintings (colour, texture, arcs, etc), dance (construct a 'grammar' for choreography), mimicking human speech patterns, etc..

[Scientific American, June 1989, Dewdney]

"I Spent an Interesting Evening Recently with a Grain of Salt"

- Mark V. Shaney

(computer-generated contributor to UseNet News group called net.singles)

- You can try it online here: http://www.yisongyue.com/shaney/

- Output of 2nd order word-level Markov Chain after training on 90,000 word philosophical essay:
- *"Perhaps only the allegory of simulation is unendurable--more cruel than Artaud's Theatre of Cruelty, which was the first to practice deterrence, abstraction, disconnection, deterritorialisation, etc.; and if it were our own past. We are witnessing the end of the negative form. But nothing separates one pole from the very swing of voting "rights" to electoral..."*

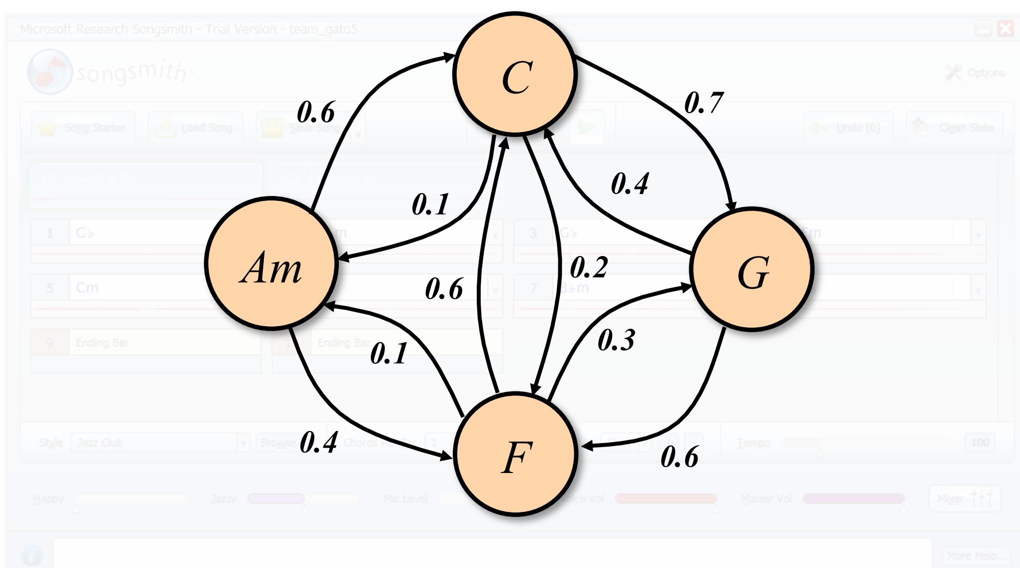
Text synthesis

- Jane Austen's *Pride and Prejudice*:
  - 121,549 words
  - 8,828 unique words (most common: 'the')
  - 7,800,000 possible pairs of words
  - 58,786 pairs (0.075%) actually appeared
  - most common pair?

  - Given a model learned from this text, we can
    - generate more "Jane Austen"-like novels
    - estimate the likelihood that a snippet of text was written by Jane Austen – this is actually a much harder and more subtle problem

- David Cope has done this for Bach chorales.

Music synthesis



- Chord progressions learned from large database of guitar tablature, but we can build far more subtle and convincing Markov models

Google's PageRank

Internet applications [edit]

The PageRank of a webpage as used by Google is defined by a Markov chain.[3] It is the probability to be at page i in the stationary distribution on the following Markov chain on all (known) webpages. If N is the number of known webpages, and a page i has k_i links then it has transition probability $\frac{\alpha}{k_i} + \frac{1-\alpha}{N}$ for all pages that are linked to and $\frac{1-\alpha}{N}$ for all pages that are not linked to. The parameter α is taken to be about 0.85.[citation needed]

Markov models have also been used to analyze web navigation behavior of users. A user's web link transition on a particular website can be modeled using first- or second-order Markov models and can be used to make predictions regarding future navigation and to personalize the web page for an individual user.

http://en.wikipedia.org/wiki/Markov_chain

Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry (1999). *The PageRank citation ranking: Bringing order to the Web.*

See also:
J. Kleinberg. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

Stationary distributions

- So why do we converge to these particular columns?

$$P^{20} = \begin{bmatrix} 0.2 & 0.44 & 0.36 \\ 0.2 & 0.44 & 0.36 \\ 0.2 & 0.44 & 0.36 \end{bmatrix}$$

- The vector [0.2 0.44 0.36] is unchanged by multiplication by $\boldsymbol{P^T}$:

$$\begin{bmatrix} 0.0 & 0.25 & 0.25 \\ 0.75 & 0.25 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{bmatrix} \begin{bmatrix} 0.2 \\ 0.44 \\ 0.36 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.44 \\ 0.36 \end{bmatrix}$$

(note the reversion to the transpose in order to perform actual matrix multiplication!)

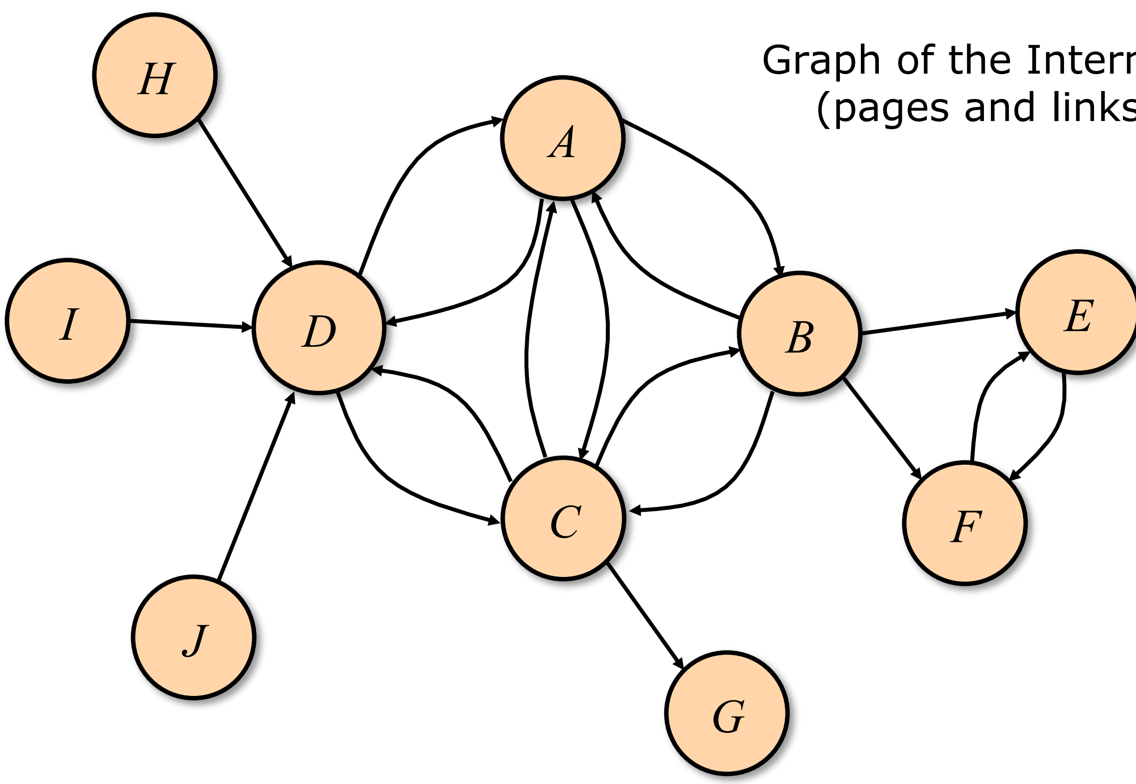
- Where have we seen this before?

Stationary distributions

- [0.2 0.44 0.36]   is an *eigenvector* of $Q = P^T$
  - a vector **x** such that $Q\,\mathbf{x} = \mathbf{x}$
  - (in linear algebra, you'll learn that the definition is a vector **x** such that $Q\,\mathbf{x} = \lambda\mathbf{x}$, so in the above, the scaling factor λ was 1; it's called the *eigenvalue*)
  - The vector **x** defines a line (all the scalar multiples of **x**), and this line is *invariant* under the action of Q
  - Such lines make for really nice coordinate axes for describing Q in the simplest possible way

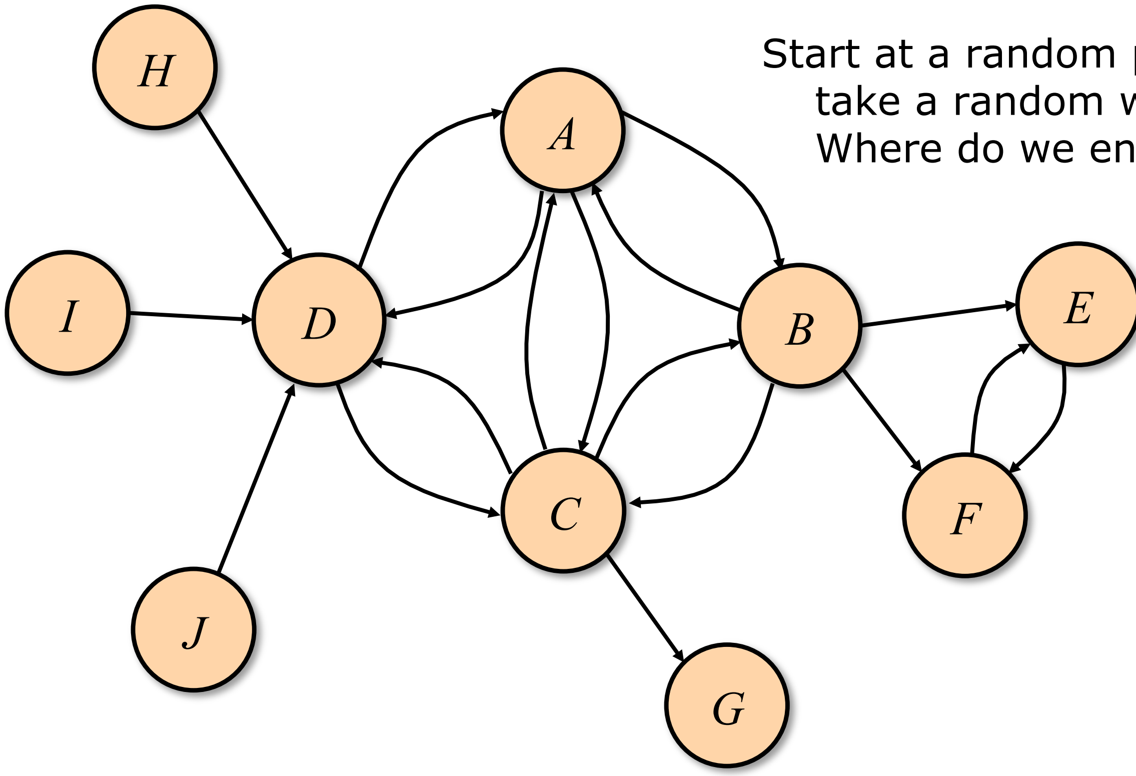
- If we look at a long sequence, this gives the proportion (spectrum) of days we expect to be nice/rainy/snowy – also a 'steady state' condition.

Google's PageRank



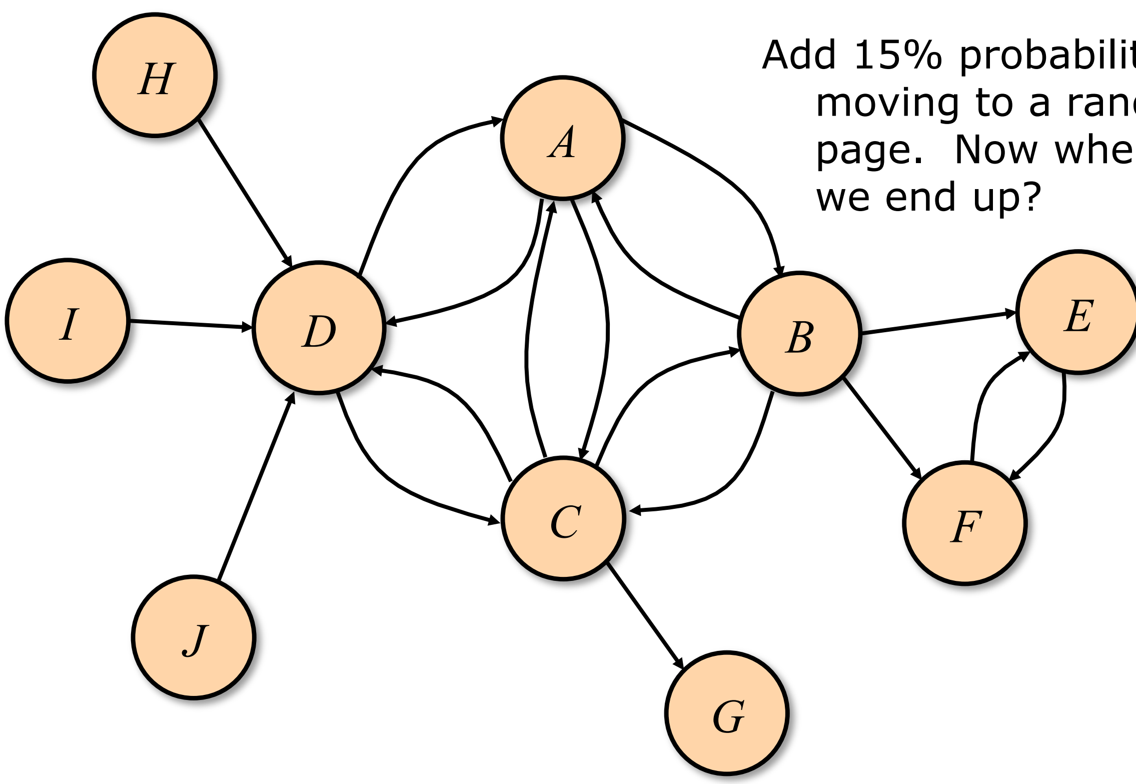
Graph of the Internet (pages and links)

Google's PageRank



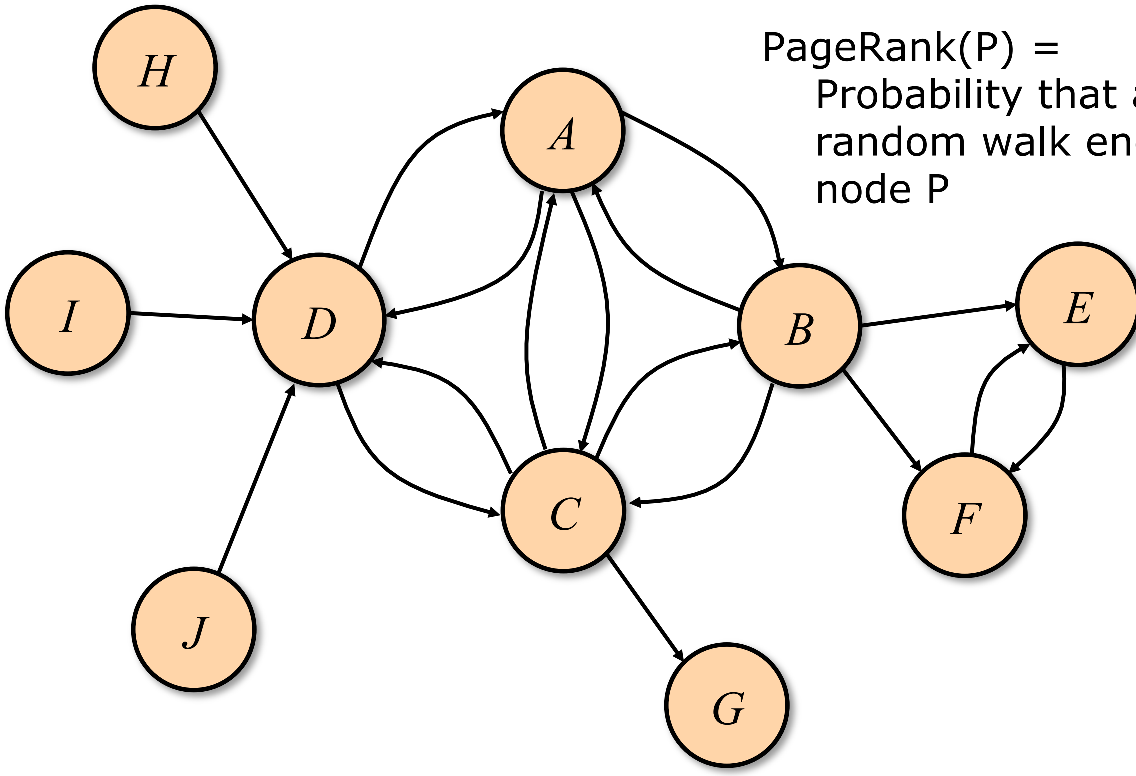
Start at a random page, take a random walk. Where do we end up?

Google's PageRank

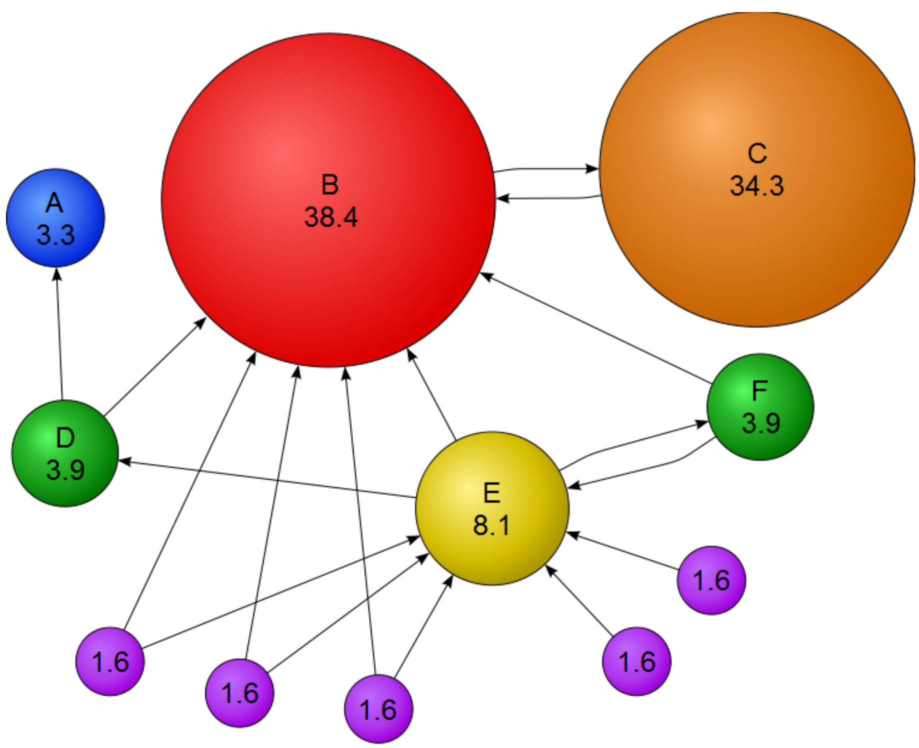


Add 15% probability of moving to a random page. Now where do we end up?

Google’s PageRank



PageRank(P) =
Probability that a long random walk ends at node P



(The ranks are an eigenvector of the transition matrix)

Back to text

- We can use Markov chains to generate new text
- Can they also help us recognize text?
  - In particular, the author?
  - Who wrote this paragraph?

    “Suppose that communal kitchen years to come perhaps. All trotting down with porringers and tommycans to be filled. Devour contents in the street. John Howard Parnell example the provost of Trinity every mother's son don't talk of your provosts and provost of Trinity women and children cabmen priests parsons fieldmarshals archbishops.”

Author recognition

- We can use Markov chains to generate new text
- Can they also help us recognize text?
  - How about this one?

    „Diess Alles schaute Zarathustra mit grosser Verwunderung; dann prüfte er jeden Einzelnen seiner Gäste mit leutseliger Neugierde, las ihre Seelen ab und wunderte sich von Neuem. Inzwischen hatten sich die Versammelten von ihren Sitzen erhoben und warteten mit Ehrfurcht, dass Zarathustra reden werde.“

The Federalist Papers

- 85 articles addressed to New York State, arguing for ratification of the Constitution (1787-1788)
- Written by "Publius" (?)
- Really written by three different authors:
  - John Jay, James Madison, Andrew Hamilton
- Who wrote which one?
  - 73 have known authors, 12 are in dispute

Author recognition

- Suppose we want to know who wrote a given paragraph/ page/article of text
- Idea: for each suspected author:
  - Download all of their works
  - Compute the transition matrix
  - Find out which author's transition matrix is the best fit to the paragraph
- What is the probability of a given *n*-length sequence *s (*aka *random walk)*?

$$s = s_1\ s_2\ s_3\ \ldots\ s_n$$

  - Probability of generating *s* = the product of transition probabilities:

$$\underbrace{\Pr(S_1 = s_1)}_{\text{Probability that a sequence starts with } s_1} \underbrace{\Pr(S_2 = s_2 | S_1 = s_1) \Pr(S_3 = s_3 | S_2 = s_2) \ldots \Pr(S_n = s_n | S_{n-1} = s_{n-1})}_{\text{Transition probabilities}}$$